Reconstruction Codes for DNA Sequences with Uniform Tandem-Duplication Errors

Yonatan Yehezkeally
Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel
yonatany@post.bgu.ac.il

Moshe Schwartz
Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel
schwartz@ee.bgu.ac.il

***Abstract*—DNA as a data storage medium has several advantages, including far greater data density compared to electronic media. We propose that schemes for data storage in the DNA of living organisms may benefit from studying the reconstruction problem, which is applicable whenever multiple reads of noisy data are available. This strategy is uniquely suited to the medium, which inherently replicates stored data in multiple distinct ways, caused by mutations. We consider noise introduced solely by uniform tandem-duplication, and utilize the relation to constant-weight integer codes in the Manhattan metric. By bounding the intersection of the cross-polytope with hyperplanes, we prove the existence of reconstruction codes with greater capacity than known error-correcting codes.**

I. Introduction

DNA is attracting considerable attention in recent years as a medium for data storage, due to its high density and longevity [5]. Data storage in DNA may provide integral memory for synthetic-biology methods, where such is required, and offer a protected medium for long-period data storage [2]. In particular, storage in the DNA of living organisms is now becoming feasible [23]; it has varied usages, including watermarking generically modified organisms [21] or research material [15], and even affords some concealment to sensitive information [6]. Naturally, therefore, data integrity in such media is of great interest.

Several recent works have studied the inherent constraints of storing and retrieving data from DNA. While desired sequences (over quaternary alphabet) may be synthesized (albeit, while suffering from substitution noise), generally data can only be read by observation of its subsequences, quite possibly an incomplete observation [16]. Moreover, the nature of DNA and current technology results in asymmetric errors which depend upon the dataset [9]. The medium itself also introduces other types of errors which are atypical in electronic storage, such as symbol/block-deletion and adjacent transpositions (possibly inverted) [10]. Finally, the purely combinatorial problem of recovering a sequence from the multiset of all its subsequences (including their numbers of incidence), was also studied [1], as well as coding schemes involving only these multisets (or their profile vectors – describing the incidence frequency of each subsequence) [22].

Other works were concerned with data storage in the DNA of a living organism. While this affords some level of protection to the data, and even propagation (through DNA replication), it is also exposed to specific noise mechanisms due to mutations. Examples of such noise include symbol insertions, deletion, substitutions (point-mutation), and duplication (including tandem- and interspersed-duplication). Therefore, schemes for data storage in live DNA must address data integrity and error-correction.

In an effort to better understand these typical noise mechanisms, their potential to generate the diversity observed in nature was studied. [11] classified the *capacity* and/or *expressiveness* of the systems of sequences over a finite alphabet generated by four distinct substring duplication rules: end-duplication, tandem-duplication, tandem-palindromic- duplication, and interspersed-duplication. Later, [12] fully characterized the expressiveness of bounded tandem-duplication systems, proved bounds on their capacity (and, in some cases, even exact values).

The generative properties of interspersed-duplication were also studied from a probabilistic point of view. [8] showed (under assumption of uniformity) that the frequencies of incidence for each subsequence converge to the same limit achieved by an i.i.d. source, thus reinforcing the notion that interspersed-duplication is–on its own–capable of generating diversity. [7] specifically looked at tandem- and end-duplication, and found exact capacities in the case of duplication length 1. It also tightly bounded the capacity of complement tandem-duplication, a process where the duplicated symbol is complemented (using binary alphabet).

Finally, error-correcting codes for data affected by tandem-duplication have been studied in [13], which presented a construction of optimal-size codes for correcting any number of errors under *uniform tandem-duplication* (fixed duplication length), computing their (and thus, the optimal-) capacity. It also presented a framework for the construction of optimal codes for the correction of a fixed number of errors. In general, it characterized the cases where the process of tandem-duplication can be traced back uniquely. Later, [18] studied the sphere-packing bound for unifrom tandem-duplication noise (as well as related error-models).

However, classical error-correction coding ignores some properties of the DNA storage channel; namely, stored information is expected to be replicated, even as it is mutated. This lends itself quite naturally to the reconstruction problem [20], which assumes that data is simultaneously transmitted

This work was supported in part by ISF grant no. 130/14.

978-1-5386-4780-6/18/$31.00©2018 IEEE

over several noisy channels, and a decoder must therefore estimate that data based on several (distinct) noisy versions of it. Solutions to this problem have been studied in several contexts. It was solved in [20] for sequence reconstruction over finite alphabets, where several error models were considered, such as substitutions, transpositions and deletions. Moreover, a framework was presented for solving the reconstruction problem in general cases of interest in coding theory, utilizing a graph representation of the error model. The problem was also studied in the context of permutation codes with transposition and reversal errors [17]. Later, applications were found in storage technologies [3], [4], since modern application might preclude the retrieval of a single data point, in favor of multiple-point requests. However, the problem hasn't been addressed yet for data storage in the DNA of living organisms, where it may be most applicable.

In this paper, we study the reconstruction problem over DNA sequences, with uniform tandem-duplication error. The paper is organized as follows: In Section II we present notations and definitions. In Section III we demonstrate that reconstruction codes are error-correcting codes and find their requisite minimal-distance, as a function of the reconstruction parameters. In Section IV we then study bounds on the sizes of such codes by an isometric embedding to constant-weight codes in the Manhattan metric. Finally, in Section V we show that by considering reconstruction codes we improve the capacity of known error-correcting codes, and conclude with closing remarks in Section VI. Throughout the paper proofs and some auxiliary propositions have been omitted, due to space restrictions; The reader is referred to the arXiv (preprint arXiv:1801.06022) for a complete version.

II. Preliminaries

Throughout this paper, though human DNA is composed of four nucleotide bases, we observe the more general case of sequences over a finite alphabet; since the alphabet elements are immaterial to our discussion, we denote it throughout as $\mathbb{Z}_q$. We observe the set of finite sequences (also: *words*) over it $\mathbb{Z}_q^* = \bigcup_{n=0}^{\infty} \mathbb{Z}_q^n$. For any two words $u, v \in \mathbb{Z}_q^*$, we denote their concatenation uv. For each word $x \in \mathbb{Z}_q^n$, we denote its *length* $|x| = n$. We also take special note of $\mathbb{Z}_q^{\geqslant k} = \{x \in \mathbb{Z}_q^* \mid |x| \geqslant k\}$. For ease of notation, we let $\mathbb{N}$ stand for the set of non-negative integers.

For $0 < k \in \mathbb{N}$, $i \in \mathbb{N}$, we define a *tandem-duplication* of *duplication-length* k by the mappings

$$\mathcal{T}_{k,i}(x) = \begin{cases} uvvw & x = uvw,\ |u| = i, |v| = k, \\ x & \text{otherwise.} \end{cases}$$

If $y = \mathcal{T}_{k,i}(x)$ and $y \neq x$ (which occurs whenever $|x| \geqslant i+k$), we say that y is a *descendant* of x, and denote $x \underset{k}{\Longrightarrow} y$. In what follows, we focus on the uniform tandem-duplication model (i.e., we fix k) because of its simplicity.

Further, given a sequence $\{x_j\}_{j=0}^{t} \subseteq \mathbb{Z}_q^*$ such that for all $0 \leqslant j < t$, $x_j \underset{k}{\Longrightarrow} x_{j+1}$, we say that x_t is a *t-descendant* of x_0, and denote $x_0 \overset{t}{\underset{k}{\Longrightarrow}} x_t$. For completeness, we also denote $x \overset{0}{\underset{k}{\Longrightarrow}} x$. Finally, if there exists some $t \in \mathbb{N}$ such that $x \overset{t}{\underset{k}{\Longrightarrow}} y$, we also denote $x \overset{*}{\underset{k}{\Longrightarrow}} y$.

We denote the set of t-descendants of $x \in \mathbb{Z}_q^*$ as

$$D_k^t(x) = \{y \in \mathbb{Z}_q^* \mid x \overset{t}{\underset{k}{\Longrightarrow}} y\},$$

for some $t \in \mathbb{N}$. We also denote the *descendant-cone* of x by $D_k^*(x) = \bigcup_{t=0}^{\infty} D_k^t(x)$.

We say that $x \in \mathbb{Z}_q^{\geqslant k}$ is *irreducible* if it is not the descendant of any word. We exclude from the definition shorter words, for which the condition vacuously holds. We denote by Irr_k the set of all irreducible words, and $\mathrm{Irr}_k(n) = \mathrm{Irr}_k \cap \mathbb{Z}_q^n$.

It was shown in [14], [19] that for each word $x \in \mathbb{Z}_q^{\geqslant k}$, a unique irreducible word exists for which x is a descendant. We call it the *root* of x, and denote it by $R_k(x)$. This induces an equivalence relation by $x \sim_k y$ if $R_k(x) = R_k(y)$.

We also follow [14] in defining, for $x \in \mathbb{Z}_q^{\geqslant k}$, $\mathrm{Pref}_k(x)$ as the length-k *prefix* of x, and $\mathrm{Suff}_k(x)$ as its suffix. Using this notation, we define an embedding $\phi_k : \mathbb{Z}_q^{\geqslant k} \to \mathbb{Z}_q^k \times \mathbb{Z}_q^*$ by

$$\phi_k(x) = \big(\mathrm{Pref}_k(x), \mathrm{Suff}_{|x|-k}(x) - \mathrm{Pref}_{|x|-k}(x)\big).$$

It is seen in [14] that this mapping is indeed injective. Further, it was shown that, defining $\zeta_{k,i} : \mathbb{Z}_q^k \times \mathbb{Z}_q^* \to \mathbb{Z}_q^k \times \mathbb{Z}_q^*$ by

$$\zeta_{k,i}(a,b) = \begin{cases} (a, b_1 0^k b_2) & b = b_1 b_2,\ |b_1| = i, \\ (a,b) & \text{otherwise,} \end{cases}$$

where $0 < k \in \mathbb{N}$, $i \in \mathbb{N}$, we have $\phi_k(\mathcal{T}_{k,i}(x)) = \zeta_{k,i}(\phi_k(x))$.

The simplicity of $\zeta_{k,i}$ in comparison to $\mathcal{T}_{k,i}$ motivates the analysis of tandem-duplications using the ϕ_k images of sequence.

If $b \in \mathbb{Z}_q^*$ is composed of the subsequences

$$b = 0^{s_1} w_1 0^{s_2} \cdots w_m 0^{s_{m+1}}; \quad w_1, \ldots, w_m \in (\mathbb{Z}_q \setminus \{0\})^*$$

we define

$$\mu(b) = 0^{s_1 \bmod k} w_1 0^{s_2 \bmod k} \cdots w_m 0^{s_{m+1} \bmod k},$$
$$\sigma(b) = \left(\left\lfloor \frac{s_1}{k} \right\rfloor, \ldots, \left\lfloor \frac{s_{m+1}}{k} \right\rfloor\right).$$

We may note that $\mathrm{wt}_H(b) = \mathrm{wt}_H(\mu(b))$, where wt_H is the Hamming weight, and $\sigma(b) \in \mathbb{N}^{\mathrm{wt}_H(b)+1} = \mathbb{N}^{\mathrm{wt}_H(\mu(b))+1}$. We also observe that b is recoverable from $\sigma(b), \mu(b)$. It was proven in [14] that if $x, y \in \mathbb{Z}_q^{\geqslant k}$, $\phi_k(x) = (a_1, b_1)$ and $\phi_k(y) = (a_2, b_2)$, then $x \sim_k y$ if and only if $a_1 = a_2$ and $\mu(b_1) = \mu(b_2)$. Moreover, $x \in \mathrm{Irr}_k$ if and only if $\sigma(b_1) = (0, 0, \ldots, 0)$. Note that, equivalently, we may say that b contains no zero-runs of length k; such sequences are called $(0, k-1)_q$-*Run-Length-Limited*, or $(0, k-1)_q$-RLL.

For $x \in \mathrm{Irr}_k$, $\phi_k(x) = (a, b)$, we denote $m(x) = \mathrm{wt}_H(b)$ and define $\psi_x : D_k^*(x) \to \mathbb{N}^{m(x)+1}$ by $\psi_x(y) = \sigma(b')$, where $\phi_k(y) = (a, b')$.

Finally, for $n \geqslant k$ and $x, y \in \mathbb{Z}_q^n$ we define

$$d_k(x,y) = \min\{t \in \mathbb{N} \mid D_k^t(x) \cap D_k^t(y) \neq \emptyset\},$$

or $d_k(x,y) = \infty$ if $\{t \in \mathbb{N} \mid D_k^t(x) \cap D_k^t(y) \neq \emptyset\} = \emptyset$. It was shown in [14, Lem. 14] that $d_k(x,y) = \infty$ if and only if

$x \not\sim_k y$, hence $d_k(\cdot,\cdot)$ is finite on $D_k^t(x)$, for any particular $x \in \mathbb{Z}_q^{\geqslant k}$. Furthermore, [14, Lem. 19] shows that for any $x \sim_k y$ with $|x| = |y|$ it holds that

$$d_k(x,y) = \tfrac{1}{2} \left\| \sigma(b_1) - \sigma(b_2) \right\|_1 ,$$

thus $d_k(\cdot,\cdot)$ defines a metric on $\mathbb{Z}_q^n$ for all $n \geqslant k$.

III. Reconstruction Codes

The reconstruction problem in the context of uniform tandem-duplication errors can be stated as follows: suppose data is encoded in $C \subseteq \mathbb{Z}_q^n$, and suppose we later are able to read distinct $x_0, x_1, \ldots, x_N \in D_k^t(c)$ for some specific $c \in C$; can we uniquely identify c?

A. Codes for reconstruction

It is apparent (see [20]) that to allow successful reconstruction we require codes to satisfy the following.

Definition 1 Take $N, t, n > 0$. We say that $C \subseteq \mathbb{Z}_q^n$ is a *uniform tandem-duplication reconstruction code*, which we abbreviate as an $(N,t,k)_q$-UTR code, if

$$\max\left\{ |D_k^t(c) \cap D_k^t(c')| \;\middle|\; c, c' \in C,\ c \neq c' \right\} \leqslant N.$$

We state this section's main result in the following corollary.

Corollary 2 Take $x \in \mathbb{Z}_q^*$, $n \geqslant |x|$. Then $C \subseteq D_k^*(x) \cap \mathbb{Z}_q^n$ is an $(N,t,k)_q$-UTR code if and only if

$$\min\{ d_k(c,c') \mid c, c' \in C,\ c \neq c' \} \geqslant d_{N,t}(m(x)),$$

where we make the notation

$$d_{N,t}(m) = \min\left\{ \delta \in \mathbb{N} \;\middle|\; \binom{t-\delta+m}{m} \leqslant N \right\}.$$

Proof: (sketch) The claim follows from a calculation of the size of intersection for two descendant-cones. ∎

B. Size of reconstruction codes

In this section we aim to estimate the maximal size of $(N,t,k)_q$-UTR codes.

Definition 3 For $m, r > 0$ we denote the $(m,r)-$*simplex*

$$\Delta_r^m = \left\{ (x_i)_{i=1}^{m+1} \in \mathbb{N}^{m+1} \;\middle|\; \sum\nolimits_{j=1}^{m+1} x_j = r \right\}.$$

Theorem 4 We take positive integers N, t and $n > k$. For $C \subseteq \mathbb{Z}_q^n$ and $x \in \mathrm{Irr}_k$ we denote $C_x = C \cap D_k^*(x)$ and define $r(x) = \frac{n-|x|}{k}$.

If $C_x \neq \emptyset$ then $r(x) \in \mathbb{N}$ and $r(x) < \left\lfloor \frac{n}{k} \right\rfloor$. Moreover, C is an $(N,t,k)_q$-UTR code if and only if for all $x \in \mathrm{Irr}_k$ such that $C_x \neq \emptyset$, the image $\psi_x(C_x) \subseteq \Delta_{r(x)}^{m(x)}$ satisfies

$$\min\left\{ \tfrac{1}{2} \left\| c - c' \right\|_1 \;\middle|\; c \neq c' \in \psi_x(C_x) \right\} \geqslant d_{N,t}(m(x)).$$

We therefore see that finding error-correcting codes for uniform tandem-duplication, after restriction to each descendant-cone, essentially amounts to finding error-correcting codes in the Manhattan metric over Δ_r^m. We start by notating the maximal size of such codes:

Definition 5 For $m, r > 0$ and $d \geqslant 0$ we define

$$M(m,r,d) = \max\left\{ |C| \;\middle|\; C \subseteq \Delta_r^m,\ \min_{\substack{c,c' \in C \\ c \neq c'}} \tfrac{1}{2} \left\| c - c' \right\|_1 \geqslant d \right\}.$$

We now note that if $C \subseteq \mathbb{Z}_q^n$, $x, x' \in \mathrm{Irr}_k(n - rk)$ (i.e., $r(x) = r(x') = r$) and $m(x) = m(x')$, then $D_k^{n-rk}(x) \cong D_k^{n-rk}(x')$ (through, e.g., $\psi_{x'}^{-1} \circ \psi_x$). It is therefore practical to assume $|C_x| = |C_{x'}| = M(m,r,d_{N,t}(m))$ for all such x, x'. This results in the following corollary, which concludes this section:

Corollary 6 If $C \subseteq \mathbb{Z}_q^n$ is an $(N,t,k)_q$-UTR code, and for all $x \in \mathrm{Irr}_k$ it holds that $|C_x| = M(m,r,d_{N,t}(m))$, then

$$|C| = \sum_{r=0}^{\lfloor n/k \rfloor - 1} \sum_{m} M(m,r,d_{N,t}(m)) \cdot q^k \cdot \cdot \left| \left\{ b \in \mathbb{Z}_q^{n-(r+1)k} \;\middle|\; \substack{b \text{ is } (0,k-1)_q\text{-RLL} \\ \mathrm{wt}_H(b) = m} \right\} \right|$$

IV. Codes on the simplex with the Manhattan metric

Corollary 6 motivates us to estimate the optimal size of error-correcting codes in the Manhattan metric over the (m,r)-simplex. This section is dedicated to that question; a key component of which will be the evaluation of the requisite minimal distance $d_{N,t}(m)$.

A. Sphere size

In this section we evaluate the size of Manhattan-metric spheres in Δ_r^m, then establish the Gilbert-Varshamov bound in the asymptotic regime.

Definition 7 For $m \in \mathbb{N}$ and $r \in \mathbb{Z}$, we denote the hyperplane

$$A_r^m = \left\{ (x_i)_{i=1}^{m+1} \in \mathbb{Z}^{m+1} \;\middle|\; \sum\nolimits_{i=1}^{m+1} x_i = r \right\}.$$

We also denote the ball of radius $d \geqslant 0$ about $x \in A_r^m$ as

$$B_r^m(d;x) = \left\{ y \in A_r^m \;\middle|\; \tfrac{1}{2} \left\| y - x \right\|_1 \leqslant d \right\};$$

since the size of each ball is invariant in r, x, we denote $B^m(d) = B_0^m(d;0)$ for ease of notation.

In the following lemma we make an asymptotic evaluation of $|B^m(d)|$:

Lemma 8 Take $\mu \in (0,1)$ and fix $d > 0$. Suppose we're given a sequence of dimensions $(m_n)_{n>0}$ such that $\lim_{n\to\infty} \frac{m_n}{n} = \mu$. Then $\lim_{n\to\infty} \frac{1}{n} \log_2 |B^{m_n}(d)| = 0$.

Using the Gilbert-Varshamov bound, we can now show the following.

Theorem 9 Take $\mu \in (0,1)$, $\rho > 0$ and integer sequences $(m_n)_{n>0}$, $(r_n)_{n>0}$ such that $\lim_{n\to\infty} \frac{m_n}{n} = \mu$ and $\lim_{n\to\infty} \frac{r_n}{n} = \rho$. Also take a fixed $d > 0$. Then

$$\lim_{n\to\infty} \tfrac{1}{n} \log_2 M(m_n, r_n, d) = (\mu+\rho) H\left(\frac{1}{1+\frac{\rho}{\mu}}\right). \quad (1)$$

B. Minimal distance of reconstruction codes

Next, given $N, t > 0$ and $m > 0$, we establish bounds on

$$d_{N,t}(m) = \min\left\{\delta \in \mathbb{N} \;\middle|\; \binom{t-\delta+m}{m} \leqslant N\right\}$$

seen in Corollary 2.

Lemma 10 If $N \leqslant m$ then $d_{N,t}(m) = t$.

V. Capacity of reconstruction codes

We can now determine the *capacity* of $(N,t,k)_q$-codes, in some asymptotic regimes. We define the *rate* of of a family of codes (C^n), $C^n \subseteq \mathbb{Z}_q^n$, as

$$R(C^n) = \limsup \tfrac{1}{n} \log_q |C^n|,$$

and the capacity of $(N,t,k)_q$-codes as the supremum of rates of families of such codes. Since [13] showed that $\mathrm{Irr}_k(n)$ can correct any number of tandem-duplication errors, they are trivially $(N,t,k)_q$-codes for all N, t. In this section we prove that reconstruction codes have strictly higher capacity.

First, we denote for any $n, r \in \mathbb{N}$ such that $n \geqslant k$ and $r < \lfloor \frac{n}{k} \rfloor$, and any $N, t \in \mathbb{N}$

$$\mathcal{M}_{N,t}(n,r) = \sum_m M(m, r, d_{N,t}(m)) \cdot \cdot \left|\left\{ b \in \mathbb{Z}_q^{n-(r+1)k} \;\middle|\; \begin{smallmatrix} b \text{ is } (0,k-1)_q\text{-RLL} \\ \mathrm{wt}_H(b) = m \end{smallmatrix} \right\}\right|.$$

We recall for all n, if $r_n = \arg\max_r \mathcal{M}_{N,t}(n,r)$, that by Corollary 6 we have an $(N,t,k)_q$-code $C \subseteq \mathbb{Z}_q^n$ with $|C| \geqslant q^k \mathcal{M}_{N,t}(n, r_n)$. Corollary 6 also implies that for all $C \subseteq \mathbb{Z}_q^n$ it holds that $|C| \leqslant \frac{n}{k} q^k \mathcal{M}_{N,t}(n, r_n)$. We therefore focus on maximizing $\limsup \frac{1}{n} \log_q \mathcal{M}_{N,t}(n, r_n)$ by choice of r_n.

In what follows, we take $\gamma \in (0,1)$ and set $r_n = \frac{1-\gamma}{k} n - 1$ for any $n \in \mathbb{N}$ for which $r_n \in \mathbb{N}$; we shall assume that such n exist, and refer only to such indinces.

For all $x \in \mathrm{Irr}_k(n - r_n k) = \mathrm{Irr}_k(k + \gamma n)$, recall that we denoted $\phi_k(x) = (a,b)$ with $b \in \mathbb{Z}_q^{\gamma n}$ in $(0,k-1)_q$-RLL. We shall build a reconstruction code in the descendant cones of only such x, which we denote C_γ.

Lemma 11 There exists a system $\mathcal{S} \subseteq (0,k-1)_q$-RLL and $\theta \in \left(\frac{1}{2}, 1\right)$ such that

$$\mathrm{cap}\, \mathcal{S} = \lim_{l\to\infty} \tfrac{1}{l} \log_q \left|\mathcal{S} \cap \mathbb{Z}_q^l\right| = \mathrm{cap}\ (0,k-1)_q\text{-RLL}$$

and for all $b \in \mathcal{S}$ it holds that $\mathrm{wt}_H(b) \geqslant \theta |b|$.

Lemma 11 implies that there exists a subset $S_k \subseteq \mathrm{Irr}_k$ such that $\mathrm{cap}\, S_k = \mathrm{cap}\, \mathrm{Irr}_k$, and for every $x \in S_k$ of length $|x| = k + \gamma n$ we have $m(x) \geqslant \lceil \theta \cdot \gamma n \rceil$. For the rest of this section we only build codes C_γ^n in the descendant cones of roots in S_k. Note, then, that if we denote $m_n = \lceil \theta \cdot \gamma n \rceil$ then the resulting codes have rate

$$R(C_\gamma^n) \geqslant \gamma\, \mathrm{cap}(\mathrm{Irr}_k) + \\ + \limsup \tfrac{1}{n} \log_q M(m_n, r_n, d_{N,t}(m_n)) \quad (2)$$

Theorem 12 As before, we denote $r_n = \frac{1-\gamma}{k} n - 1$ and $m_n = \lceil \theta \cdot \gamma n \rceil$. Then, assuming t is fixed and $N_n = o(n)$, and if we denote $\mathcal{H}(x) = xH\left(\frac{1}{x}\right)$ for $x \geqslant 1$,

$$\lim_n \tfrac{1}{n} \log_q M(m_n, r_n, d_{N_n,t}(m_n)) = \\ = \frac{\theta\gamma}{\log_2 q} \cdot \mathcal{H}\left(1 + \frac{1-\gamma}{k\theta\gamma}\right)$$

Proof: (sketch) By Lemma 10 we have that $d_{N_n,t}(m_n)$ is fixed. We note that $\lim_{n\to\infty} \frac{r_n}{n} = \frac{1-\gamma}{k}$ and $\lim_{n\to\infty} \frac{m_n}{n} = \theta\gamma$, hence by Theorem 9 the claim is proven. ■

It's worth noting that, in practical applications, we may indeed expect the number of duplications t, which is dependent on the period of time before data is read, to be fixed w.r.t. n. The allowed uncertainty N_n will also likely be bounded, which Theorem 12 accommodates.

Before moving on to show that generally $R(C_\gamma^n)$ may be made to exceed $\mathrm{cap}(\mathrm{Irr}_k)$ by a careful choice of γ, we look at the following example.

Example 13 Set $q = k = 2$. Then the Perron eigenvalue of $T_2(1)$ is $\lambda = \frac{1+\sqrt{5}}{2}$, and

$$\mathrm{cap}(\mathrm{Irr}_2) = \log_2(\lambda) = \log_2\left(\frac{1+\sqrt{5}}{2}\right) \approx 0.6942.$$

In addition, any θ which is less than $\pi_1 = \frac{1}{2}\left(1 + \frac{1}{\sqrt{5}}\right) \approx 0.7236$ satisfies Lemma 11.

Alternatively, we may set $q = 4$ (for the special case of human DNA) and duplication-length $k = 2$. Now the Perron eigenvalue of $T_4(1)$ is given by $\lambda = \frac{3+\sqrt{21}}{2}$, hence

$$\mathrm{cap}(\mathrm{Irr}_2) = \log_4(\lambda) = \log_4\left(\frac{3+\sqrt{21}}{2}\right) \approx 0.9613.$$

Further, we may choose any θ which is less than $\pi_1 = \frac{1}{2}\left(1 + \sqrt{\frac{3}{7}}\right) \approx 0.8273$.

$R(C_\gamma^n)$ is shown as a function of $\gamma \in (0,1)$ for both cases in Figure 1, under the assumptions of asymptotic regime made in Theorem 12. The figure demonstrates that the capacity of reconstruction codes (bounded from below by the maximum of the curve) is greater than $\mathrm{cap}(\mathrm{Irr}_k)$. □

We now attempt to maximize $R(C_\gamma^n)$ by a proper choice of γ. Theorem 12 motivates the following definition:

Definition 14 Take $\mathcal{R}, \theta \in \left(\frac{1}{2}, 1\right)$. We define $R : (0,1) \to \mathbb{R}$ by

$$R(\gamma) = \gamma\mathcal{R} + \frac{\theta\gamma}{\log_2 q} \mathcal{H}\left(1 + \frac{1-\gamma}{k\theta\gamma}\right).$$

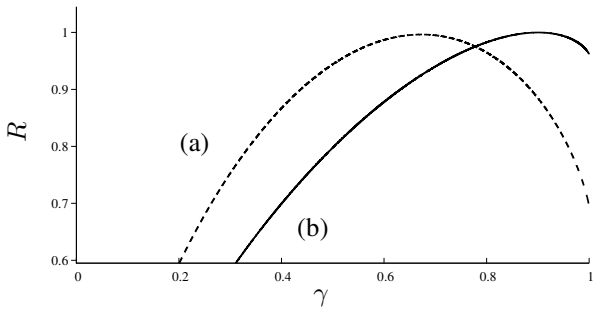

Figure 1. Rate $R(C_\gamma^n)$ in the cases (a) $q = k = 2$, $\theta = 0.7236$, and (b) $q = 4$, $k = 2$, $\theta = 0.8273$. The value at $\gamma = 1$ equals $\mathrm{cap}(\mathrm{Irr}_k)$, which is the capacity of known error-correcting codes ($N = 0$).

Analysis of $R(\gamma)$ is simpler using the following change of variable:

Definition 15 Define $x : (0,1) \to (0,\infty)$ by $x(\gamma) = \frac{1-\gamma}{\gamma}$.

We observe that x is a decreasing diffeomorphism, and $\gamma = \frac{1}{1+x(\gamma)}$.

We can now show that there always exists a choice of γ for which we get $R(C_\gamma^n) > \mathrm{cap}(\mathrm{Irr}_k)$:

Theorem 16 $\max_{\gamma\in(0,1)} R(\gamma) > \mathcal{R}$.

Proof: (sketch) Observe that $R(\gamma)$ is continuously differentiable and satisfies $\lim_{\gamma\to 0} R(\gamma) = 0$, $\lim_{\gamma\to 1} R(\gamma) = \mathcal{R}$. We can show that $R'(\gamma) = 0$ if and only if

$$q^{-k\mathcal{R}} = \left(1 + \frac{x(\gamma)}{k\theta}\right)^{k\theta-1} \cdot \frac{x(\gamma)}{k\theta} \tag{3}$$

This equation has a unique solution $x_0 = x(\gamma_0)$, since the RHS is a monotonic increasing function of x, vanishing at $x = 0$ and unbounded as x grows. Moreover, $0 < x_0 < k\theta$, since $k\theta > 1$, hence the RHS is greater than 1 at $x = k\theta$. Thus $R(\gamma)$ has a unique local extremum in $(0,1)$.

It now suffices to show that $R(\gamma)$ is concave, which may be verified by examining the second derivative. Hence, the extremum is a maximum, as claimed. ∎

VI. Conclusion

We have proposed that reconstruction codes can be applied to data-storage in the DNA of living organisms, due to the channel's inherent property of data replication.

We have showed, under the assumption of uniform tandem-duplication noise, that reconstruction codes are error-correcting codes with minimal distance dependent on the reconstruction parameters. We then proved the existence of such codes with rates surpassing that of know error-correcting codes.

We believe that further research should focus on explicit code constructions. It is also desirable to examine the problem under broader noise models, such as bounded tandem-duplication, interspersed-duplication (perhaps inversed), as well as combinations of multiple error models.

References

[1] J. Acharya, H. Das, O. Milenkovic, A. Orlitsky, and S. Pan, "String reconstruction from substring compositions," *SIAM J. Discrete Math.*, vol. 29, no. 3, pp. 1340–1371, 2015.

[2] F. Balado, "Capacity of DNA data embedding under substitution mutations," *IEEE Trans. on Inform. Theory*, vol. 59, no. 2, pp. 928–941, Feb. 2013.

[3] Y. Cassuto and M. Blaum, "Codes for symbol-pair read channels," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 8011–8020, Dec. 2011.

[4] Y. M. Chee, H. M. Kiah, A. Vardy, V. K. Vu, and E. Yaakobi, "Coding for racetrack memories," *IEEE Trans. on Inform. Theory*, 2018.

[5] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[6] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," *Nature*, vol. 399, no. 6736, pp. 533–534, 1999.

[7] O. Elishco, F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of some Pólya string models," in *Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT2016), Barcelona, Spain*, Jul. 2016, pp. 270–274.

[8] F. Farnoud, M. Schwartz, and J. Bruck, "A stochastic model for genomic interspersed duplication," in *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT2015), Hong Kong, China*, Jun. 2015, pp. 904–908.

[9] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *IEEE Trans. on Inform. Theory*, vol. 63, no. 8, pp. 4982–4995, Aug. 2017.

[10] R. Gabrys, E. Yaakobi, and O. Milenkovic, "Codes in the Damerau distance for DNA storage," in *Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT2016), Barcelona, Spain*, Jul. 2016, pp. 2644–2648.

[11] F. F. Hassanzadeh, M. Schwartz, and J. Bruck, "The capacity of string-duplication systems," *IEEE Trans. on Inform. Theory*, vol. 62, no. 2, pp. 811–824, Feb. 2016.

[12] S. Jain, F. F. Hassanzadeh, and J. Bruck, "Capacity and expressiveness of genomic tandem duplication," *IEEE Trans. on Inform. Theory*, vol. 63, no. 10, pp. 6129–6138, Oct. 2017.

[13] S. Jain, F. F. Hassanzadeh, M. Schwartz, and J. Bruck, "Duplication-correcting codes for data storage in the DNA of living organisms," *IEEE Trans. on Inform. Theory*, vol. 63, no. 8, pp. 4996–5010, Aug. 2017.

[14] ——, "Noise and uncertainty in string-duplication systems," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT2017), Aachen, Germany*, Jun. 2017, pp. 3120–3124.

[15] D. C. Jupiter, T. A. Ficht, J. Samuel, Q.-M. Qin, and P. de Figueiredo, "DNA watermarking of infectious agents: Progress and prospects," *PLoS Pathog*, vol. 6, no. 6, p. e1000950, 2010.

[16] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. on Inform. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[17] E. Konstantinova, "On reconstruction of signed permutations distorted by reversal errors," *Discrete Math.*, vol. 308, pp. 974–984, 2008.

[18] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Bounds on codes correcting tandem and palindromic duplications," *arXiv preprint arXiv:1707.00052*, 2017.

[19] P. Leupold, C. Martín-Vide, and V. Mitrana, "Uniformly bounded duplication languages," *Discrete Appl. Math.*, vol. 146, no. 3, pp. 301–310, 2005.

[20] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. on Inform. Theory*, vol. 47, no. 1, pp. 2–22, Jan. 2001.

[21] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leiherer, and R. Wagner, "Embedding permanent watermarks in synthetic genes," *PLoS ONE*, vol. 7, no. 8, p. e42465, 2012.

[22] N. Raviv, M. Schwartz, and E. Yaakobi, "Rank modulation codes for DNA storage," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT2017), Aachen, Germany*, Jun. 2017, pp. 3125–3129.

[23] S. L. Shipman, J. Nivala, J. D. Macklis, and G. M. Church, "CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria," *Nature*, vol. 547, p. 345, Jul. 2017.